11-7-2017

Number of Predictors and Multicollinearity: What are their Effects on Error and Bias in Regression?

Matthew Ryan Lavery
Bowling Green State University, mlavery@bgsu.edu

Parul Acharya
Columbus State University

Stephen A. Sivo
University of Central Florida

Lihua Xu
University of Central Florida

Number of Predictors and Multicollinearity:

What Are Their Effects on Error and Bias in Regression?

Multicollinearity, or collinearity, is often encountered in applied regression models, and occurs when two or more predictor variables are correlated. Multicollinearity becomes a problematic condition when it influences the inferences made about significance and parameter estimates. High levels of multicollinearity may lead to large variances in the least squares estimators of beta coefficients in the regression equation. It may also have unpredictable and inconsistent effects on parameter estimates and significance and may lead to biased results. Multicollinearity masks the true relationship of the predictor variables with the dependent variable, thereby undermining the unique variance explained by predictors in the model.

The magnitude of correlation between the predictor variables has an important influence on the variance inflation factor (VIF). VIF is the reciprocal of tolerance (Dormann et al., 2013):

$$VIF_k = \frac{1}{1 - \left(\rho_{X_k X_{\neg k}}\right)^2}, \quad (1)$$

Tolerance is 1 minus the proportion of variance a predictor variable shares with the other predictor variables in the regression model (shown here as the squared correlation between the predictor in question, X_k, and the other predictors in the model, $X_{\neg k}$). Tolerance represents the proportion of variance in a predictor variable that is not shared or related to the other predictor variables. In the literature, a number of rules or criteria have been recommended to indicate when VIF or tolerance values are considered to be very high to the extent that it may bias the regression results. Sometimes a VIF of 10 is considered to be excessive or a VIF as low as 4 have been used to indicate high levels of multicollinearity between the predictor variables. It is

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

important to understand the extent to which predictor variables are correlated with each other and with the outcome variable in the context of VIF and other factors that influence the variance of regression coefficients and the bias in the model. Kraha, Turner, Nimon, Zientek and Henson (2012) explained the role that predictor variables play in making the regression coefficients in conjunction with other factors such as correlation coefficients, β weights, and structure coefficients in order to interpret and correct for collinearity in light of theoretical and statistical significance.

Previous research has already established that the correlations between predictor variables influences VIF and may introduce bias in the regression results (Azen, Budescu, Reiser, 2001; Johnson, 2001; Johnson & Lebreton, 2004; Kraha, et al., 2012). However, there are no specific guidelines for researchers to understand how the number of predictors, magnitude of inter-correlations between them and proportion of variance jointly explained in the outcome variable by the predictor variables influence the amount of VIF and bias in the regression model. In the present study, we systematically varied the sample size and analyzed parameter bias, model bias, rates of Type I and Type II error, and VIF values produced under various multicollinearity conditions with two, four, and six predictors. The objective of the present study is to provide specific guidance to applied researchers regarding the degree of multicollinearity that can be problematic for multiple regression depending on the number of predictors modeled and the degree of inter-correlations between them.

Review of the Literature

Multicollinearity is a common phenomenon that occurs in regression when two or more predictors are correlated with one other and commonly occurs in almost all regression-based

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

procedures. Multicollinearity is a nuisance condition which impacts model development, estimation and interpretation especially when predictors share notably strong correlations. For instance, multicollinearity can make it difficult to parse out the unique contribution of each predictor variable towards explaining the dependent variable variance. Predictors are assumed to be non-collinear in a regression design (Lomax, 2007; Stevens, 2007). Hence, multicollinearity makes it difficult to evaluate the individual importance of each predictor in a model. Adding to this, high levels of multicollinearity do not allow unique estimates of the regression coefficients associated with predictors because the coefficients are interchangeable, making it difficult to assess the distinctive relationship between a predictor and the dependent variable relative to the relationship between another predictor and that same dependent variable (Gunst & Mason 1980; Marquardt & Snee 1975). As collinearity increases, the standard errors of the regression coefficients also increase making the coefficients less stable across samples and less representative of the population-level estimates. So multicollinearity not only affects the evaluation of predictor contributions but also makes regression coefficients unreliable (Gunst & Mason 1980; Marquardt & Snee 1975; O'Brien, 2007) because the coefficients are more likely to vary from sample to sample. This means the regression equation formed will be unstable across samples as well (Kutner, Nachtsheim, Neter, & Li, 2005). Related to this, Mason and Perreault (1991) conducted a Monte Carlo experiment which suggested the effect of multicollinearity on coefficient standard errors is a problem only when multicollinearity among predictors are extreme.

Mela and Kopalle (2002) studied the effects of collinearity on the correlations, parameter inference, variable omission bias and diagnostic indices in regression. The authors suggested that

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

positive and negative correlations of equal magnitude can have different effects on the inflation of parameter variance estimation and that one correlation structure can severely influence variable omission bias and have little effect on variance inflation whereas other correlation structure can have the opposite effects. So, to determine the consequences of multicollinearity it is important to examine an array of collinearity diagnostics - not just one. Using this strategy, Mela and Kopalle (2002) found that positive correlations between predictor variables yields less precise estimates, can influence coefficients to change signs and affect effect size of the model. However, negative correlations among predictor variables have greater influence on variable omission bias than equivalent positive correlations

The degree of multicollinearity may be assessed either through careful analysis of a correlation matrix of the predictors used in the model, or by examining the variance inflation factor (VIF) for each predictor. In the literature, the VIF has been a common method of identifying multicollinearity. As the name suggests, VIF indicates how much of the variance in the dependent variable explained by the predictor variables is inflated. In other words, the VIF is directly related to the regression coefficient associated with a predictor variable, and it provides a clear assessment of the influence of collinearity on the estimated variance of the regression coefficient (O'Brien, 2007). A VIF of 8 suggests that (keeping all other factors constant) the variance of the regression coefficient associated with a particular predictor variable would be 8 times greater than it would have been if the predictor variable had been linearly independent with the other predictor variables in the regression model. The VIF indicates the extent to which a predictor variable has a strong linear relationship with the other predictor variable(s) in an ordinary least square regression analysis and accounts for the inflation in the estimated

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

regression coefficients when the predictor variables share a high correlation between them or the lack of independence between them. Though the VIF is easily produced in most statistical analysis software and is often consulted by applied researchers, the literature offers unclear guidance about which values for VIF are too large and how the researcher should respond (O'Brien, 2007).

O'Brien (2007) provided recommendations to cope with the issue of multicollinearity. One way to deal with the issue of collinearity is to combine predictor variables that are conceptually similar and share high correlation into a single measure and then utilize the newly created measure in the regression model which would take care of collinearity created by the high correlation between the two variables and would usually provide a more reliable estimate of the variable of interest. Another way is to remove the collinear variable from the model. There are some general guidelines that have been followed in the literature for assessing when VIF is a cause of concern:

- If VIF is greater than 10 then the predictor variables are highly correlated indicating high levels of multicollinearity and is a matter of concern (Bowerman & O'Connell, 1990).
- If the average VIF is substantially greater than 1, then the regression coefficients and the models may be biased (Bowerman & O'Connell, 1990).

Mason and Perreault (1991) examined how multicollinearity influences Type II error rates in addition to the accuracy of regression coefficients and associated standard errors. They varied the degree of collinearity, the values of true regression coefficients, sample size, and the model R^2. However, they did not vary the number of predictor variables and kept it fixed at four. The results suggested that collinearity should be viewed along with power and its interaction

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

with the sample size, the values of true regression coefficients and the overall fit of the model. The simulations results suggest that small sample size and low R^2 interacts with high collinearity to produces inaccurate regression coefficients. The results also suggested that Type II error is high when sample size is small or the overall fit of the model is low and any combination of these factors along with high levels of collinearity makes it difficult to obtain reliable inferences.

Blaze and Ye (2012) studied the effects of multicollinearity on the parameter estimates and standard errors in multilevel models by designing a Monte Carlo simulation study in which they included a two-level predictor model with correlation between level-1 and level-2 predictors and group-mean centering level-1 predictors. They varied the intra-class correlation coefficients, number of groups and cases per group. Their simulation findings were consistent with other simulation studies in the literature examining effects of multicollinearity in regression analysis. High levels of multicollinearity inflated the standard errors and the estimate of the intercept for the random slope component was biased when multicollinearity existed between level-1 predictors. The fixed effects remained relatively stable even at high levels of multicollinearity. There was an increase in positive bias of standard error estimates with increase in inter-class correlation coefficient.

The current simulation study adds to the regression literature on understanding the influence of predictors, multicollinearity, VIF and bias in several different meaningful ways. First, the study considers the effects of multicollinearity with both Type I and Type II error rates on parameter estimation bias, model bias and reported VIF values. Previous studies have not simultaneously examined the effects of different types of bias and VIF on parameters. Second, research has demonstrated that sample size is an important factor to consider when evaluating the

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

robustness of regression results (Frank & Friedman, 1993; Jean, Kerneis & Porcher, 2008; Greenland, Sander, Schwartzbaum & Finkle, 2000). Kiers and Smilde (2007) found that the ordinary least square regression yielded best regression model results with sample size of 10, 20, and 50 data points. Kroll and Song (2013) studied the performance of four regression techniques (ordinary least squares, variance inflation factor screening, principal component regression, and partial least squares regression) to evaluate the impact of multicollinearity (by varying sample sizes, inter-correlations between predictor variables and model error variances) on different regression models. They found that the undesirable influences of multicollinearity are magnified at smaller sample sizes, higher correlations between predictor variables and larger unexplained or error variances. In the current study, the sample size was systematically varied according to the rules of thumb by Green (1991). Third, the number of predictor variables were varied to two, four and six. Fourth, the proportion of variance in the outcome variable explained by the predictor variables was systematically varied according to small, medium and large according to Cohen's (1992) recommendations and was then used to calculate the model bias. Fifth, how much is the model and parameter bias when these factors interact with VIF. Sixth, the behavior of VIF was examined as a function of the correlations between the predictor variables, correlation between each predictor and outcome variable (proportion of variance explained; PVE), sample size, error rates (Type I and Type II), bias (model and parameter), and how the coefficients and standard errors change when these factors interact with VIF.

Methods and Data

The present study uses a Monte Carlo simulation design (Fan, Felsovalyi, Sivo, & Keenan, 2002) in which data is generated with known properties, analyzed using least squares

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

multiple regression, and the results are analyzed to determine the effects of multicollinearity and the number of predictors in the model on significance, parameter estimates, and the VIF. Data sets are generated for analysis with one of the following regression equations:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon\,, \tag{2.a}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon\,, \tag{2.b}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon\,. \tag{2.c}$$

In all data sets, the outcome variable is Y ~N(0, 1). Data sets generated for Equation (2.a) had two predictor variables, X_1 and X_2; data sets for equation (2.b) had four predictors, X_1 through X_4; and data sets for Equation (2.c) had six predictors, X_1 through X_6. In all cases, X_k ~N(0, 1). Predictors were generated with four multicollinearity conditions, one in which the predictors are uncorrelated ($\rho_{X_k X_{\neg k}} = 0$), and correlations of .3, .6, and .9, as well. Predictors were generated to share .1, .3, and .5 of the outcome variable's variance, resulting in a small, medium, and large proportion of variance explained (PVE), respectively (Cohen, 1992).

Simulated data sets were initially generated to reflect the rules of thumb recommended by Green (1991), where small samples contained $n = 50 + 8m$ observations and medium samples contained $n = 104 + m$ observations (where m = number of predictors), and where large samples contained $n = 500$ observations. The rules of thumb suggested by Green were used for the small and medium samples because of their prevalence in the literature. Preliminary findings indicated that the initial small sample sizes lacked the necessary statistical power (discussed in Results), thus the researchers used Green's (1991) recommended medium sample of $n = 104 + m$ observations as the small sample in this study, using $n = 250$ and $n = 500$ as the medium and large samples, respectively.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Samples were then analyzed using PROC REG in SAS 9.4. For each of the 108 experimental conditions, data was generated and analyzed 1000 times and the rates of Type I and Type II error were calculated. Model and parameter bias was also calculated as the difference between the parameter estimates produced by the regression procedure and the known value as a proportion of the known theoretical value. Model bias was calculated as

$$Bias_{model} = \frac{(AdjR^2 - PVE)}{PVE}, \quad (3)$$

where *PVE* is the known proportion of variance in the dependent variable jointly explained by the predictors which was used during data generation. Parameter bias was calculated as

$$Bias_{parameter} = \frac{\beta_{X_k} - \rho_{X_kY}}{\rho_{X_kY}}, \quad (4)$$

where ρ_{X_kY} is the known correlation between the predictor and the dependent variable, also used during data generation. These methods of calculating bias produce a value that indicates the degree to which the estimate is inflated or deflated as a proportion of the known value, where negative values indicate that the parameter has been underestimated and positive values indicate that the parameter has been overestimated. When *Bias* = 0, the estimate produced by the regression procedure matches the known value exactly. The Variance Inflation Factor (VIF) was also collected and analyzed across all iterations.

Results

Type I and Type II Error Rates

Results indicate that neither multicollinearity nor number of predictors has an effect on the Type I error rate. Across all iterations and all conditions, the mean rate of Type I error for overall model significance was .049 (SD = .007), and the mean rate of Type I error for individual

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

predictors was .049 (SD = .005). Both at the model and predictor levels, the observed Type I error rate was appropriate for the selected alpha of .05. The Type II error rate, however, was influenced both by multicollinearity and number of predictors (see Table 1).

According to Green (1991), a small sample of $n = 50 + 8m$ should provide sufficient power to detect a medium effect for the model as a whole, but not necessarily for its individual predictors, while a medium sample of $n = 104 + m$ should be sufficient to evaluate the individual predictors, as well. The preliminary findings suggested that the rules of thumb offered by Green are insufficient to provide adequate power. The results in Table 1 indicate that Green's recommended medium sample (used as the small sample in the present study) only provides adequate power to determine model significance when detecting a large effect (PVE = .5; Cohen, 1992), but never performs well enough to correctly evaluate predictor significance under the conditions simulated.

Researchers conducted logistic regression analysis to assess the relative contributions of simulated conditions to the prediction of making a Type II error both for the model and for the individual predictors in the model using the equation

$$P(TypeII) = \frac{1}{1 + e^{-\left(b_0 + b_1 PVE + b_2 \rho_{X_k X_{\neg k}} + b_3 m + b_4 S + b_5 L + b_6 \left(m \times \rho_{X_k X_{\neg k}}\right)\right)}}. \quad (5)$$

In Equation (5), PVE is the known proportion of variance in Y jointly explained by the predictors in the model; $\rho_{X_k X_{\neg k}}$ is the known collinearity among predictors defined as the correlation between a given predictor (X_k) and all other predictors in the model ($X_{\neg k}$); m is the number of predictors in the model; S and L are dummy-coded variables for small and large samples, respectively, comparing them to the un-coded medium samples; and $m \times \rho_{X_k X_{\neg k}}$ represents the

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

interaction between number of predictors and collinearity. The full model was significantly better at predicting Type II error for model *F*-tests than a constant-only model, $\chi^2(6, N = 108{,}000) = 72{,}035.03$, $p < .001$, increasing prediction accuracy from 61.6% to 84.9%. The full model was also significantly better at predicting Type II error for individual predictor *t*-tests than a constant-only model, $\chi^2(6, N = 432{,}000) = 106{,}827.05$, $p < .001$, increasing prediction accuracy from 79.3% to 85.2%. Logistic regression results for both model and predictor Type II error rates are shown in Table 2.

As the proportion of the dependent variable's variance explained by the predictors increases, the Type II error rate decreases substantially for both the regression model's *F*-tests, Exp(B) < .001, and for individual predictors' *t*-tests, Exp(B) = .002. Increased collinearity and additional predictors both increase the odds ratio of Type II errors, with Exp(B) = 1.267 and Exp(B) = 1.907 respectively for models, and with Exp(B) = 7.086 and Exp(B) = 1.853 respectively for individual predictors. The interaction between increased collinearity and additional predictors appears to have a mitigating effect on Type II error, however, as the interaction term lowers the odds ratio with Exp(B) = .524 for models and Exp(B) = .890 for predictors.

Model and Parameter Bias

Simulation results indicate that both collinearity and number of predictors in the model are related to bias. The results in Table 3 demonstrate that, across all simulated conditions, the proportion of variance explained by the model was consistently underestimated by the regression procedure, *M* = -.816, *SD* = .131. Researchers used the linear regression shown in Equation (6),

$$Bias = \beta_0 + \beta_1 PVE + \beta_2 \rho_{X_k X_{\neg k}} + \beta_3 m + \beta_4 S + \beta_5 L + \beta_6 \left(m \times \rho_{X_k X_{\neg k}}\right) + \varepsilon\,, \qquad (6)$$

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

to analyze the degree to which simulated conditions explain observed model bias and parameter bias. The regression model was significant for models, $F(6, 107993) = 11934.97$, $p < .001$, explaining 39.9% of the variance in model bias. Dummy-coded variables for small and large sample sizes did not significantly contribute to the explanation of model bias, however, with $t(107993) = -1.608$, $p = .108$, and $t(107993) = .145$, $p = .884$, respectively. All other variables included in the model were significant at $p < .001$ and are listed in Table 4. Equation (6) also significantly explained 3.5% of the variance in predictor bias, with $F(6, 431993) = 2622.19$, $p < .001$. The proportion of variance explained (PVE) did not contribute significantly to parameter bias, $t(431993) = -.225$, $p = .822$. Dummy-coded variables for small and large sample sizes also did not significantly contribute to the explanation of parameter bias, $t(431993) = -.172$, $p = .863$, and $t(431993) = .225$, $p = .822$, respectively. All other simulation conditions significantly explained parameter bias, $p < .001$, and are listed in Table 4.

In order to better understand the influence of the various conditions simulated on bias, researchers analyzed the mean and standard deviation of model and parameter bias values for each of the thirteen simulated conditions. Inspection of the values in Table 5 suggests that, as the proportion of variance in the dependent variable jointly explained by the predictors in the model increases, underestimation of the model's adjusted R-squared is decreased (from $M = -.938$ to $M = -.694$) while the variability in mis-estimation of predictor coefficients is reduced (from $SD = 2.390$ to $SD = .540$). As collinearity increases, underestimation of the model's adjusted R-squared decreases from $M = -.909$ to $M = -.723$ while the variability in model bias simultaneously increases (from $SD = .145$ to $SD = .220$). By contrast, underestimation of predictor parameters steadily increases as collinearity increases (from

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

M = -.005 to *M* = -.732), while no clear pattern emerges from the standard deviations of parameter bias values. As the number of predictors in the model increase, underestimation and variability in bias simultaneously increase both for the model's adjusted R-squared (from *M* = -.785, *SD* = .186, to *M* = -.839, *SD* = .203) as well as for predictor coefficients (from *M* = -.273, *SD* = .877, to *M* = -.544, *SD* = 1.731), though the problem appears more severe for predictor estimates. Likewise, as sample sizes increase underestimation and variability in bias simultaneously decrease both for the model's adjusted R-squared (from *M* = -.817, *SD* = .250, to *M* = -.815, *SD* = .150) as well as for predictor coefficients (from *M* = -.472, *SD* = 2.009, to *M* = -.470, *SD* = .956), though the reduction in the variability of parameter estimates is most pronounced.

Variance Inflation Factor (VIF)

Simulation results indicate that both collinearity and number of predictors in the model are related to VIF. The results shown in Table 6 indicate that VIF values reported by the regression procedure are greatest when collinearity is greatest, and that reported VIFs are highest within collinearity conditions when more predictors are included in the model. Interestingly, none of the observed VIF values approached 10, the commonly used rule of thumb for indication of problematic collinearity (O'brien, 2007). In order to analyze the extent to which simulation conditions explain the observed VIF values, researchers conducted the linear regression

$$VIF = \beta_0 + \beta_1 PVE + \beta_2 \rho_{X_k X_{\neg k}} + \beta_3 m + \beta_4 S + \beta_5 L + \beta_6 \left(m \times \rho_{X_k X_{\neg k}}\right) + \varepsilon\,. \qquad (7)$$

The regression model was significant, $F(6, 431993) = 165082.58$, $p < .001$, explaining 69.6% of the variance in parameter VIF. While $\rho_{X_k X_{\neg k}}$ has, perhaps unsurprisingly, the highest estimated coefficient in the explanation of VIF (see Table 7), examination of the standardized

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

beta-weights indicates that the interaction between collinearity and number of predictors may be more influential than collinearity alone. In fact, while increasing the number of predictors may explain a slight reduction in the observed VIF, the interaction term overshadows this reduction by the magnitude of the increase in VIF that it explains. Only the proportion of variance in the dependent variable jointly explained by the predictors (*PVE*) did not significantly explain the variance in observed VIF values, $t(431993) = -.482$, $p = .630$. All other parameters were significant at $p < .001$.

Discussion

The present study investigates how the number of predictors in a multiple linear regression model interacts with collinearity to affect the rates of Type I and Type II error, parameter estimation bias, and reported VIF values. Although the findings indicate that the number of predictors is unrelated to the Type I error rate, it affects every other outcome assessed in this study. Increasing the number of predictors by one nearly doubles the odds of making a Type II error under the conditions simulated. The number of predictors in the model inflated the odds of Type II error in determining the significance of the regression model more than any other condition simulated except for small samples of $n = 104 + m$. Collinearity inflated the odds of Type II errors in determining predictor significance more than three times as much as did the number of predictors, though the interaction between number of predictors and collinearity reduced the odds of Type II error somewhat, mitigating the problem of error inflation slightly.

The number of predictors included in a regression model was also significantly related to the underestimation of both the proportion of variance in the dependent variable explained by the regression model as well as in the beta weights of the predictors included in the regression

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

model. The interaction between number of predictors and collinearity behaves differently in the calculation of the regression model's adjusted R^2 than it does in the estimation of predictors' beta weights, however. In determining the proportion of variance explained by the model, the interaction between collinearity and number of predictors increases bias (or, more accurately, decreases the magnitude of underestimation) by the same amount that the number of predictors increase the magnitude of underestimation on its own. Thus, in the presence of collinearity, increasing the number of predictors simultaneously increases and decreases model bias, greatly mitigating its effect. When estimating predictor beta weights, however, the number of predictors, collinearity, and the interaction between number of predictors and collinearity all decrease bias (again, more accurately, increase the magnitude of parameter underestimation), greatly compounding the problems of collinearity.

In the absence of collinearity, the number of predictors included in the model has no noticeable effect on VIF. In fact, according to the results of the linear regression analyses reported in Table 7, the number of predictors slightly reduces the VIF values reported for each predictor. In the presence of collinearity, however, VIF values inflate substantially. Increased VIF values can be explained primarily by collinearity and by the interaction between number of predictors and collinearity. In this case, it would be more appropriate to say that the number of predictors, *m*, slightly mitigates the VIF inflation caused by collinearity and the interaction effect, rather than the interaction effect mitigating the two main effects.

Interestingly, comparison of the means and standard deviations of predictor bias across simulated conditions (Table 5) indicates that, as the number of predictors' increases, the standard deviation of parameter bias increases, as well. This suggests that, while across repeated

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

regression analyses, one could expect the beta weights of predictors to be underestimated by a greater magnitude than comparable models with fewer predictors, there is no way to be certain whether the beta weights for any single analysis are underestimated. The increased variance in parameter bias for models with more predictors makes the results of any single regression even more unpredictable than collinearity, itself. Taken together, the results of the present study emphasize the importance of parsimony, particularly when predictors are correlated.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

References

Azen, R., Budescu, D. V., & Reiser, B. (2001). Criticality of predictors in multiple regression. *British Journal of Mathematical and Statistical Psychology*, *54*, 201-225.

Blaze, T. J., & Ye, F. (2012). The Effect of Within- and Cross-Level Multicollinearity on Parameter Estimates and Standard Errors in Multilevel Modeling with Different Centering Methods. Paper Presented at American Educational Research Association Conference, Vancouver, British Columbia, Canada.

Biau, D. J., Kerneis, S., & Porcher, R (2008). Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research. *Clinical Orthopaedics and Related Research*, *466*(9), 2282-2288.

Bowerman, B. L., & O.Connell, R. T. (1990). *Linear Statistical Models: An Applied Approach* (2nd ed.), Belmont, CA: Duxbury.

Chaterjee, S.; Hadi, A.S. (2006). Regression analysis by example. 4th.ed. New York: John Wiley, 408pp.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering, 14*(3), 391-403. doi: 10.1081/QEN-120001878

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., ... Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2002). *SAS® for Monte Carlo studies: A guide for quantitative researchers*. Cary, N.C: SAS Institute.

Ganzach, Y. (1998). Nonlinearity, multicollinearity and the probability of Type II error in detecting interaction. *Journal of Management, 24*(5), 615-622.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*, 499 - 510.

Greenland, S., Schwartzbaum, J. A., & Finkle, W. D. (2000). Problems due to Small Samples and Sparse Data in Conditional Logistic Regression Analysis. *American Journal of Epidemiology, 151*(5), 531-539.

Gunst, R. F., & Mason, R. L. (1980). *Regression analysis and its application: A data-oriented approach*. New York: Marcel Dekker.

Johnson J. W. (2001). Determining the relative importance of predictors in multiple regression: Practical applications of relative weights. *Advances in Psychology Research*, Vol. V, 231–251.

Johnson, J. W., & Lebreton, J. M. (2004). History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods, 7*(3), 238-257.

Kiers, H., and A. Smilde (2007). A comparison of various methods for multivariate regression with highly collinear variables, *Statistical Methods Application*, *16*(2), 193–228.

Kraha, A., Turner, H., Nimon, K., Zientek., & Henson, R. K. (2012). Tools to Support Interpreting Multiple Regression in the Face of Multicollinearity. *Frontiers in Psychology, 3*(44).

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Kroll, C. N., & Song, P. (2013). Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research*, *49*(6), 3756-3769.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). New York, NY: McGraw-Hill/Irwin.

Lauridsen, J., & Mur, J. (2006). Multicollinearity in cross-sectional regression. *Journal of Geographical Systems, 8,* 317-333.

Lomax, R. G. (2007). *An introduction to statistical concepts* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

Mason, C. H., & Perreault, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research, 28,* 268-280.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician, 29,* 3-20.

Mela, C. F., & Kopalle, P. K (2002). The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations, *Applied Economics*, *34*(6), 667-677.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), 673-690. doi: 10.1007/s11135-006-9018-6.

Robinson, C., & Schumacker, R. E. (2009). Interaction effects: Centering, variance inflation factor, and interpretation issues. *Multiple Linear Regression Viewpoints, 35*(1), 6-11.

Stevens, J. P. (2007). *Intermediate statistics: A modern approach* (3rd ed.). New York, NY: Routledge.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Table 1

Type II Error Rate for the Model (Predictors in Parentheses) by Known Proportion of Variance Explained (PVE) and Collinearity Conditions

Sample	*m*	*PVE* = .1	*PVE* = .3	*PVE* = .5	Sample	*m*	*PVE* = .1	*PVE* = .3	*PVE* = .5
		$\rho_{X_k X_{\neg k}} = 0$					$\rho_{X_k X_{\neg k}} = .3$		
Small	2	.915 (.924)	.487 (.641)	.060 (.224)	Small	2	.901 (.931)	.397 (.686)	.024 (.266)
	4	.935 (.945)	.810 (.875)	.460 (.734)		4	.905 (.939)	.627 (.889)	.146 (.764)
	6	.946 (.949)	.884 (.919)	.694 (.860)		6	.921 (.946)	.716 (.921)	.288 (.880)
Medium	2	.847 (.880)	.146 (.340)	.001 (.013)	Medium	2	.795 (.874)	.055 (.370)	.000 (.019)
	4	.920 (.934)	.556 (.776)	.103 (.493)		4	.872 (.936)	.243 (.803)	.002 (.522)
	6	.939 (.943)	.741 (.880)	.359 (.746)		6	.887 (.938)	.368 (.890)	.006 (.763)
Large	2	.724 (.789)	.004 (.057)	.000 (.000)	Large	2	.665 (.812)	.000 (.096)	.000 (.001)
	4	.867 (.910)	.208 (.611)	.003 (.194)		4	.790 (.919)	.021 (.660)	.000 (.228)
	6	.918 (.934)	.492 (.796)	.041 (.522)		6	.866 (.944)	.075 (.829)	.000 (.581)
		$\rho_{X_k X_{\neg k}} = .6$					$\rho_{X_k X_{\neg k}} = .9$		
Small	2	.864 (.924)	.305 (.761)	.008 (.393)	Small	2	.866 (.948)	.221 (.894)	.000 (.752)
	4	.905 (.942)	.480 (.910)	.034 (.827)		4	.884 (.944)	.340 (.945)	.003 (.912)
	6	.913 (.952)	.589 (.937)	.065 (.901)		6	.915 (.953)	.388 (.948)	.007 (.938)
Medium	2	.776 (.907)	.023 (.495)	.000 (.053)	Medium	2	.721 (.932)	.006 (.807)	.000 (.500)
	4	.854 (.941)	.088 (.857)	.000 (.662)		4	.771 (.946)	.028 (.931)	.000 (.875)
	6	.861 (.943)	.163 (.918)	.000 (.838)		6	.841 (.952)	.044 (.935)	.000 (.921)
Large	2	.582 (.863)	.002 (.200)	.000 (.000)	Large	2	.557 (.927)	.000 (.670)	.000 (.221)
	4	.711 (.931)	.000 (.772)	.000 (.413)		4	.626 (.947)	.000 (.904)	.000 (.788)
	6	.772 (.943)	.004 (.874)	.000 (.721)		6	.672 (.944)	.001 (.926)	.000 (.896)

Note. "Small" samples contained $n = 104 + m$ observations (where m = number of predictors), "Medium" samples contained 250 observations, and "Large" samples contained 500 observations for each simulation run. Model error rates are calculated for $n = 1000$ runs per simulated condition. Predictor error rates include all predictors included in 1000 simulation runs per condition (i.e., $n = 2000$ for each $m = 2$ condition, $n = 4000$ for each $m = 4$ condition, and $n = 6000$ for each $m = 6$ condition). Significance of model *F*-tests and predictor *t*-tests were determined at $p < .05$.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Table 2

Estimates for Logistic Regression of Type II Error for Regression Models and Predictors

	Model			Predictors		
Predictor	B(SE)	Wald	Exp(B)	B(SE)	Wald	Exp(B)
b_0	1.709 (.045)	1428.14	5.525	.151 (.022)	45.30	1.163
PVE	-15.109 (.092)	26894.56	.000	-6.041 (.031)	38706.82	.002
$\rho_{X_k X_{\neg k}}$	.236 (.075)*	10.02	1.267	1.958 (.039)	2575.50	7.086
m	.646 (.010)	4004.52	1.907	.617 (.005)	18194.52	1.853
S	1.538 (.024)	3972.47	4.656	.703 (.012)	3647.86	2.021
L	-.916 (.024)	1502.404	.400	-.663 (.010)	4271.23	.515
$m * \rho_{X_k X_{\neg k}}$	-.645 (.018)	1343.65	.524	-.116 (.009)	180.46	.890

Note. n = 108,000 for logistic regression of model error and n = 432,000 for logistic regression of predictor error. b_0 is the logistic regression constant term, PVE is the known proportion of variance in the dependent variable jointly explained by the predictors in the model, $\rho_{X_k X_{\neg k}}$ is collinearity expressed as the known correlation between a given predictor (k) and all other predictors ($\neg k$) in the model, m is the number of predictors in the regression model, S and L are dummy coded variables comparing small samples (n = 104 + m; Green, 1991) and large samples (n = 500), respectively, to un-coded medium samples (n = 250), and $m * \rho_{X_k X_{\neg k}}$ is an interaction term for number of predictors and collinearity.

* $p < .01$; all other parameters $p < .001$

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Table 3

Theoretical Bias for the Model (Predictors in Parentheses) by Known Proportion of Variance Explained (PVE) and Collinearity Conditions

Sample	*m*	*PVE* = .1	*PVE* = .3	*PVE* = .5	Sample	*m*	*PVE* = .1	*PVE* = .3	*PVE* = .5
		$\rho_{X_k X_{\neg k}} = 0$					$\rho_{X_k X_{\neg k}} = .3$		
Small	2	-.960 (-.079)	-.844 (.034)	-.754 (-.002)	Small	2	-.941 (-.256)	-.813 (-.239)	-.684 (-.238)
	4	-.972 (.023)	-.926 (-.005)	-.873 (.011)		4	-.937 (-.463)	-.859 (-.474)	-.768 (-.475)
	6	-.984 (-.014)	-.952 (-.026)	-.918 (-.014)		6	-.954 (-.591)	-.870 (-.594)	-.797 (-.601)
Medium	2	-.950 (-.005)	-.853 (-.011)	-.755 (-.008)	Medium	2	-.927 (-.201)	-.806 (-.231)	-.676 (-.231)
	4	-.976 (-.002)	-.926 (.012)	-.879 (-.014)		4	-.951 (-.447)	-.857 (-.472)	-.765 (-.475)
	6	-.988 (-.045)	-.947 (.000)	-.920 (-.013)		6	-.949 (-.589)	-.871 (-.592)	-.791 (-.600)
Large	2	-.948 (.001)	-.846 (.013)	-.753 (-.006)	Large	2	-.937 (-.240)	-.809 (-.237)	-.676 (-.231)
	4	-.973 (.026)	-.925 (.004)	-.877 (-.007)		4	-.952 (-.480)	-.859 (-.474)	-.764 (-.475)
	6	-.985 (.006)	-.950 (.003)	-.917 (.002)		6	-.963 (-.617)	-.876 (-.603)	-.790 (-.599)
		$\rho_{X_k X_{\neg k}} = .6$					$\rho_{X_k X_{\neg k}} = .9$		
Small	2	-.908 (-.336)	-.771 (-.385)	-.611 (-.382)	Small	2	-.907 (-.488)	-.718 (-.474)	-.527 (-.474)
	4	-.929 (-.646)	-.794 (-.649)	-.655 (-.645)		4	-.898 (-.731)	-.735 (-.733)	-.544 (-.731)
	6	-.938 (-.746)	-.813 (-.756)	-.678 (-.753)		6	-.922 (-.818)	-.725 (-.817)	-.548 (-.818)
Medium	2	-.923 (-.385)	-.758 (-.372)	-.604 (-.377)	Medium	2	-.902 (-.466)	-.724 (-.481)	-.530 (-.476)
	4	-.939 (-.660)	-.791 (-.643)	-.652 (-.643)		4	-.898 (-.726)	-.729 (-.733)	-.543 (-.731)
	6	-.926 (-.746)	-.802 (-.750)	-.668 (-.750)		6	-.911 (-.818)	-.725 (-.819)	-.547 (-.819)
Large	2	-.922 (-.376)	-.762 (-.377)	-.603 (-.377)	Large	2	-.910 (-.484)	-.716 (-.474)	-.532 (-.477)
	4	-.928 (-.640)	-.789 (-.641)	-.648 (-.642)		4	-.909 (-.733)	-.724 (-.730)	-.537 (-.729)
	6	-.934 (-.750)	-.801 (-.751)	-.670 (-.751)		6	-.905 (-.815)	-.723 (-.818)	-.541 (-.818)

Note. "Small" samples contained $n = 104 + m$ observations (where m = number of predictors; Green, 1991), "Medium" samples contained 250 observations, and "Large" samples contained 500 observations. Values reported in the table are means of $n = 1000$ simulation runs per condition. Mean predictor bias include all predictors from 1000 simulation runs per condition (i.e., $n = 2000$ for each $m = 2$ condition, $n = 4000$ for each $m = 4$ condition, and $n = 6000$ for each $m = 6$ condition). Model bias = $\frac{(AdjR^2 - PVE)}{PVE}$ and predictor bias = $\frac{\beta_{X_k} - \rho_{X_k Y}}{\rho_{X_k Y}}$ where PVE and $\rho_{X_k Y}$ are known.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Table 4

Linear Regression Estimates for the Explanation of Model Bias and Parameter Bias

	Model			Predictors		
Predictor	*b*(SE)	*t*	*p*	*b*(SE)	*t*	*p*
β_0	-.992 (.002)	-429.283	< .001	.017 (.013)	1.259	.208
PVE	.608 (.003)	214.814	< .001	-.003 (.014)	-.225	.822
$\rho_{X_k X_{\neg k}}$	.107 (.004)	29.248	< .001	-.432 (.022)	-19.751	< .001
m	-.025 (.000)	-52.037	< .001	-.029 (.003)	-11.660	< .001
S	-.002 (.001)	-1.608	.108	-.001 (.005)	-.172	.863
L	.000 (.001)	.145	.884	.001 (.005)	.225	.822
$m * \rho_{X_k X_{\neg k}}$	.025 (.001)	29.412	< .001	-.074 (.004)	-16.659	< .001

Note. n = 108,000 for linear regression of model bias and n = 432,000 for linear regression of predictor bias. β_0 is the linear regression constant term, PVE is the known proportion of variance in the dependent variable *Y* jointly explained by the predictors in the model, $\rho_{X_k X_{\neg k}}$ is collinearity expressed as the known correlation between a given predictor (k) and all other predictors ($\neg k$) in the model, m is the number of predictors in the regression model, S and L are dummy coded variables comparing small samples ($n = 104 + m$; Green, 1991) and large samples (n = 500), respectively, to un-coded medium samples (n = 250), and $m * \rho_{X_k X_{\neg k}}$ is an interaction term for number of predictors and collinearity. Model bias = $\frac{(AdjR^2 - PVE)}{PVE}$ and predictor bias = $\frac{\beta_{X_k} - \rho_{X_k Y}}{\rho_{X_k Y}}$.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Table 5

Comparison of Mean and Standard Deviation of Model Bias and Predictor Bias across Simulated Conditions

Simulated Conditions	Model			Predictors		
	n	*M*	*SD*	*n*	*M*	*SD*
All Conditions	108,000	-.816	.196	432,000	-.471	1.494
PVE = .1	3600	-.938	.205	144,000	-.471	2.390
= .3	3600	-.816	.137	144,000	-.471	.834
= .5	3600	-.694	.157	144,000	-.472	.540
$\rho_{X_k X_{\neg k}}$ = 0	2700	-.909	.145	108,000	-.005	2.188
= .3	2700	-.846	.169	108,000	-.495	1.090
= .6	2700	-.786	.192	108,000	-.693	.918
= .9	2700	-.723	.220	108,000	-.732	1.340
m = 2	3600	-.785	.186	72,000	-.273	.877
= 4	3600	-.824	.195	144,000	-.460	1.340
= 6	3600	-.839	.203	216,000	-.544	1.731
n = 104 + *m*	3600	-.817	.250	144,000	-.472	2.009
n = 250	3600	-.816	.174	144,000	-.471	1.322
n = 500	3600	-.815	.150	144,000	-.470	.956

Note. PVE is the known proportion of variance in the dependent variable *Y* jointly explained by the predictors in the model, $\rho_{X_k X_{\neg k}}$ is collinearity expressed as the known correlation between a given predictor (k) and all other predictors ($\neg k$) in the model, *m* is the number of predictors in the regression model. Model bias = $\frac{(AdjR^2 - PVE)}{PVE}$ and predictor bias = $\frac{\beta_{X_k} - \rho_{X_k Y}}{\rho_{X_k Y}}$.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Table 6

Mean VIF of Predictors (Standard Deviation in Parentheses) by Known Proportion of Variance Explained (PVE) and Collinearity Conditions

Sample	*m*	*PVE* = .1	*PVE* = .3	*PVE* = .5	Sample	*m*	*PVE* = .1	*PVE* = .3	*PVE* = .5
		$\rho_{X_k X_{\neg k}} = 0$					$\rho_{X_k X_{\neg k}} = .3$		
Small	2	1.010 (0.015)	1.010 (0.015)	1.010 (0.014)	Small	2	1.111 (0.068)	1.115 (0.066)	1.108 (0.066)
	4	1.029 (0.024)	1.030 (0.024)	1.030 (0.025)		4	1.238 (0.103)	1.243 (0.108)	1.237 (0.104)
	6	1.049 (0.032)	1.048 (0.031)	1.048 (0.031)		6	1.320 (0.122)	1.325 (0.124)	1.320 (0.126)
Medium	2	1.004 (0.006)	1.004 (0.006)	1.004 (0.005)	Medium	2	1.105 (0.043)	1.104 (0.045)	1.103 (0.045)
	4	1.013 (0.011)	1.012 (0.010)	1.012 (0.010)		4	1.220 (0.067)	1.222 (0.067)	1.219 (0.064)
	6	1.021 (0.013)	1.021 (0.013)	1.021 (0.013)		6	1.288 (0.076)	1.282 (0.077)	1.286 (0.075)
Large	2	1.002 (0.003)	1.002 (0.003)	1.002 (0.003)	Large	2	1.102 (0.030)	1.101 (0.030)	1.102 (0.029)
	4	1.006 (0.005)	1.006 (0.005)	1.006 (0.005)		4	1.210 (0.046)	1.209 (0.045)	1.212 (0.045)
	6	1.010 (0.007)	1.010 (0.006)	1.010 (0.006)		6	1.272 (0.052)	1.272 (0.052)	1.271 (0.053)
		$\rho_{X_k X_{\neg k}} = .6$					$\rho_{X_k X_{\neg k}} = .9$		
Small	2	1.592 (0.191)	1.597 (0.195)	1.583 (0.190)	Small	2	5.427 (0.962)	5.462 (1.000)	5.395 (0.986)
	4	2.038 (0.287)	2.047 (0.289)	2.046 (0.289)		4	7.954 (1.523)	7.899 (1.432)	7.920 (1.448)
	6	2.253 (0.333)	2.252 (0.326)	2.246 (0.326)		6	8.924 (1.591)	8.845 (1.619)	8.912 (1.621)
Medium	2	1.575 (0.124)	1.577 (0.124)	1.574 (0.120)	Medium	2	5.338 (0.626)	5.320 (0.618)	5.306 (0.607)
	4	1.998 (0.181)	1.991 (0.180)	2.003 (0.180)		4	7.708 (0.923)	7.681 (0.912)	7.667 (0.917)
	6	2.180 (0.207)	2.180 (0.204)	2.182 (0.204)		6	8.604 (1.042)	8.596 (1.027)	8.611 (1.031)
Large	2	1.569 (0.083)	1.564 (0.085)	1.565 (0.089)	Large	2	5.278 (0.437)	5.276 (0.417)	5.291 (0.428)
	4	1.985 (0.126)	1.980 (0.121)	1.988 (0.123)		4	7.624 (0.644)	7.669 (0.661)	7.643 (0.636)
	6	2.141 (0.140)	2.152 (0.140)	2.156 (0.142)		6	8.492 (0.695)	8.486 (0.704)	8.472 (0.725)

Note. "Small" samples contained $n = 104 + m$ observations (where m = number of predictors; Green, 1991), "Medium" samples contained 250 observations, and "Large" samples contained 500 observations. Values reported in the table are calculated from the output of 1000 simulation runs per condition, thus $n = 2000$ for each $m = 2$ condition, $n = 4000$ for each $m = 4$ condition, and $n = 6000$ for each $m = 6$ condition.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750

Table 7

Linear Regression Estimates for the Explanation of Parameter VIF

Predictor	*b*(SE)	β	*t*	*p*
β_0	.372 (.015)	—	25.643	< .001
PVE	-.007 (.015)	.000	-.482	.630
$\rho_{X_k X_{\neg k}}$	3.403 (.024)	.393	142.668	< .001
m	-.115 (.003)	-.059	-42.171	< .001
S	.085 (.006)	.014	14.308	< .001
L	-.030 (.006)	-.005	-5.047	< .001
$m * \rho_{X_k X_{\neg k}}$	.781 (.005)	.478	160.493	< .001

Note. n = 432,000. β_0 is the linear regression constant term, PVE is the known proportion of variance in the dependent variable *Y* jointly explained by the predictors in the model, $\rho_{X_k X_{\neg k}}$ is collinearity expressed as the known correlation between a given predictor (k) and all other predictors ($\neg k$) in the model, *m* is the number of predictors in the regression model, *S* and *L* are dummy coded variables comparing small samples ($n = 104 + m$; Green, 1991) and large samples (n = 500), respectively, to un-coded medium samples (n = 250), and $m * \rho_{X_k X_{\neg k}}$ is an interaction term for number of predictors and collinearity.

The Version of Record of this manuscript has been published and is available in *Communications in Statistics - Simulation and Computation*, 07 Nov 2017, doi:10.1080/03610918.2017.1371750